Maori-English Machine Translation

Mahsa Mohaghegh
Unitec
Department of Computing
Auckland, New Zealand
mmohaghegh@unitec.ac.nz

Michael McCauley
University of Auckland
Faculty of Engineering
Auckland, New Zealand
mmcc362@aucklanduni.ac.nz

Mehdi Mohammadi
Western Michigan University
Department of Computer Science
MI, USA
mehdi.mohammadi@wmich.edu

ABSTRACT

In this paper we present our research over machine translation for Maori to English language pairs. The aim of this research is to create a lookup table for Maori-English words or phrases that are extracted from a set of aligned sentences. A major problem in this domain is word/phrase alignment. To overcome this problem, some common approaches have been investigated including statistical analyzing as well as heuristic methods. Although the research is in its initial steps and needs more parallel data, the initial results show that we can rely on statistical approaches to create Maori-English machine translation.

Categories and Subject Descriptors

G.3 [Probability and Statistics]: Statistical computing; I.2.7 [**Artificial Inteligence**]: Natural Language Processing – *language models, machine translation.*

General Terms

Algorithms, Languages.

Keywords

Machine Translation, Statistical Machine Translation, Maori language.

1. INTRODUCTION

As the main application of Natural Language Processing (NLP), Machine Translation (MT) is becoming a necessary tool in nowadays rapid and voluminous stream of digital content. This need could be more sensible by increasing cross-regional communication as well as information exchange. For example, many TV channels broadcast with closed caption to different nations who have different languages. Or, some communities like European Union require documents to be translated to several languages simultaneously.

Considering the technology used in Machine Translation (MT) systems, we could categorize MT systems into two different types: rule-based and empirical approaches. The translation process in the rule-based approaches is defined and conducted by a set of rules developed manually by linguistic experts.

In the core of empirical approaches, we acquire necessary knowledge to perform the translation by automatic analysis of translation examples. In this approach that is also referred as corpus-based approach, a MT system for new languages and domains could be developed as quickly as providing sufficient training data. By success of MT system in one domain, the system could be expanded to other domains in a language pair. A common architecture of an empirical machine translation system is shown in Figure 1.

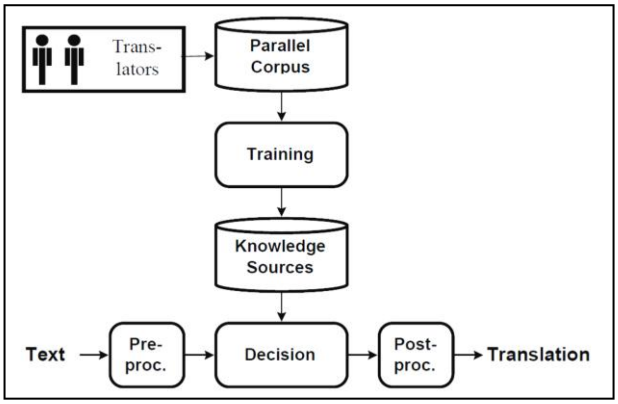

Figure 1.Common architecture of corpus-based machine translation systems [1]

Empirical MT systems also could be developed by two different approaches: Example-based MT (EBMT) and Statistical MT (SMT). To translate a new sentence in EBMT approach, similar translation examples to that sentence that have been seen previously are analyzed. In Statistical approaches, a statistical translation model is computed automatically from the translation examples.

Among the aforementioned approaches, SMT is the dominant methodology for generating machine translation systems in recent years and it is attracting more attention of researchers towards its improvement and development [2]. However, one active problem in this field is word alignment of bilingual training data. Word alignment could simply be defined as mapping source language words to their corresponding translation words in the target language. In a professional view, a word alignment is an applicable hidden parameter in Statistical Machine Translation [3].

The problem of word alignment in a bilingual sentence-aligned corpus is addressed in many works. There are some common approaches in this field in which statistical methods are widely used.

Based on Och and Ney [3], a general definition of alignment between two word strings of source and target languages could be represented by a subset of the Cartesian product of the word positions in both word strings. However, because of difficulty in implementation of such general models, most of alignment models are restricted in some ways. One typical approach is One-to-One alignment [4] in a sentence pair ($F= f_1 \ldots f_I$, $E= e_1 \ldots e_J$) in which I and J are the length of the source and target sentences in terms of words, respectively. The alignment A would be represented as a subset of $\{1,2, ..., I\}\times\{1,2, ...,J\}$. A source word in the position i is mapped to a word of target language in position j, if $(i, j) \in A$. Mappings in this model may contain assignment to an empty string in the target language. Even though this model could be used to build a bilingual vocabulary automatically, but it may not yield good results for a phrase-based SMT purpose.

In this work, our parallel corpus is aligned at sentence level. By the alignment approaches we try to fine-grain the alignment granularity to the phrase and word level.

2. RELATED WORKS

The statistical alignment models make the base of statistical translation models and were initially word based. IBM Models 1-5

[5], HMM [6] and Model 6 [3] are some remarkable instances of this category. One successful implementation of IBM alignment models is Giza++ that generates good quality alignments although the size of parallel corpus is rather small [4].

However, in many language pairs and in Maori-English as well, co-occurrences of multi-words[1] are prevalent. Therefore, having relationship between multi-words in an alignment process is indispensable [7]. Hence, decomposing the source and target sentences to only separated words could not lead to a good alignment.

But, emerge of phrase-base models made significant advances in the field of SMT. Specially, recent research has exploited the strength of syntax-based approaches to make hierarchical phrase-based model for SMT [8]. This model tries to find a translation for segments of words (phrases) using synchronous context-free grammars defined as rules.

A wide range of word alignment systems have been used to facilitate creating dictionaries automatically. Uplug [9] is a system that has been used in some research projects like [10-13]. Uplug is a Perl script including a set of language processing modules such as word alignment, sentence alignment, POS-tagging, term extraction from parallel corpora, etc. Its word alignment process finds word alignment candidates using a combination of several statistical parameters. It also contains GIZA++ in its standard package to align words and phrases.

In order to describe links between multi-word units, Tiedemann [7] proposed combination of single word links. The author investigated different clue alignment strategies using this approach. The clue alignment method exploits a way of combination of association indicators on a word-to-word level. The result of combination would be a two-dimensional clue matrix. Data elements of this matrix express the collected evidence of an association between word pairs in bi-text segments taken from a parallel corpus. Word alignment is then the task of identifying the best links according to the associations indicated in the clue matrix.

Word alignment models can be categorized as generative and discriminative models [14]. Well-known IBM models [5] and the HMM model are generative models. On the other hand, manual alignments in word alignment tasks result in discriminative models.

3. WORD ALIGNMENT MODELS

Word alignment of a bilingual aligned corpus is a core task of SMT and has been addressed in many research applications. At the main part of these approaches, a model of the translation process should be created in which the word alignment is a hidden variable. Along with statistical models, some heuristic models like Dice coefficient are also exploited. Computation of word alignments at these approaches are based on analyzing some association score of a link between the words of source language and target language.

The following is an example of alignment and the correspondence between the source and target words (Figure 2):

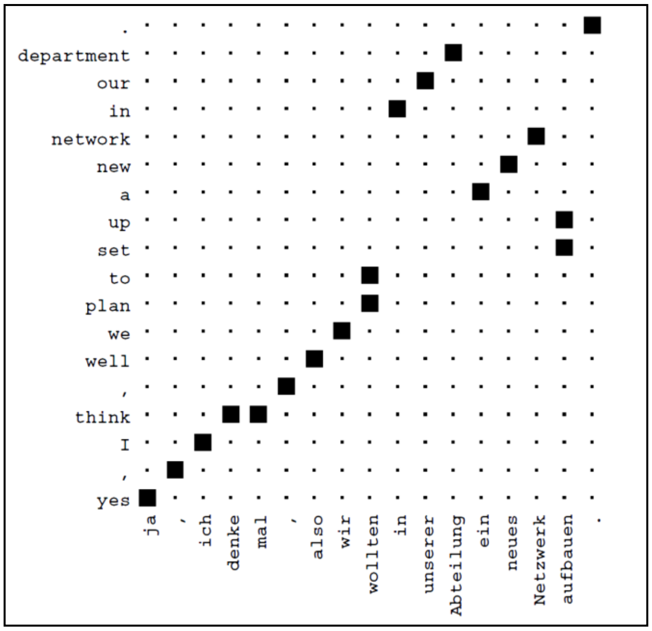

Figure 2.An example of word alignment [1]

Each word mapping shows an association $j \rightarrow i = a_j$ in which the alignment is between the source position j to the target position $i = a_j$. The alignment mappings may have some association of $a_j=0$ to indicate that there is no aligned words in the target language for the source word. Here e_0 is a symbol of empty word in target language.

3.1 Statistical Alignment Models

Having a source language sentence f_1^J and a target language sentence e_1^I, to model the relationship between the source sentence and the target one in statistical machine translation, we rely on the translation probability $Pr(f_1^J| e_1^I)$. At the aforementioned model, a hidden parameter $a=a_1^J$ is introduced that lead us to the alignment model $Pr(f_1^J, a_1^J | e_1^I)$. This parameter reveals an association from a source position j to a target position a_j. The translation model and the alignment model are related based on the following equation:

$$Pr(f_1^J|e_1^I) = \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) \qquad (1)$$

Statistical model usually is affected by some unknown parameters υ which revealed by learning from the training data. The dependency of the model to the parameters could be stated as the following equation:

$$Pr(f_1^J, a_1^J|e_1^I) = p_\theta(f_1^J, a_1^J|e_1^I) \qquad (2)$$

Using a parallel corpus consisting of S sentence pairs, we could perform the training of unknown parameters υ. These parameters are identified by likelihood maximization over the training corpus:

$$\hat{\theta} = \underset{\theta}{argmax} \left\{ \prod_{s=1}^{S} \left[\sum_{a} p_\theta(\mathbf{f}_s, \mathbf{a}|\mathbf{e}_s) \right] \right\} \qquad (3)$$

To figure out this maximization in statistical models, one useful tool is the EM algorithm [1]. There may be several alignments for a sentence pair, but the best alignment is always the desired one by:

[1] Multi-word: A group of words that their combination makes an expression that is not predictable from the meaning of individual words.

$$\hat{a}_1^J = \underset{a_1^J}{argmax} \; p_{\hat{\theta}}(f_1^J, a_1^J | e_1^I) \qquad (4)$$

Expectation Maximization (EM) [15] is a commonly used statistical inference model for word alignment problem [16]. In order to acquire alignment distribution, EM just considers the most likely word connections in the parameter space and ignores the other less likely contributions [17]. At the first step of EM algorithm, we build all possible connections between words of each sentence pair. The point here is that all connections are equally likely. Then we learn from the corpus that some connections occur more frequently. So, the inference would be more frequent connection, more likely alignment. After calculating all connection probabilities, the structure hidden in the parallel corpus will be revealed by EM algorithm and all source words will be aligned to their counterparts in the target language.

3.2 Heuristic Models

In these models, a simple method for extracting word alignments is used based on a similarity measurement between the units of text of the two languages. In many cases, the Dice coefficient is used for similarity measurement. All possible association between the words of the source sentence and those of the target sentence and their score are constructed:

$$dice(e_i, f_j) = \frac{2 \cdot C(e_i, f_j)}{C(e_i) \cdot C(f_j)} \qquad (5)$$

At the above equation, *C(e)* shows the number of occurrences of word *e* in the target sentences and *C(f)* is associated to the count of words *f* in the source sentences. *C(e, f)* represents the co-occurrence count of word *e* and word *f* in the parallel corpus. Here, the word alignment could be determined using the largest score:

$$a_j = \underset{i}{\mathrm{argmax}}\{dice(e_i, f_j)\} \qquad (6)$$

[18] reports another version of this approach called competitive linking algorithm in which after aligning highest score associations, these alignment are eliminated from the alignment matrix until every word in the source language or those in the target language is aligned.

Heuristic models in contrast to statistical models are simple to develop as well as easy to understand. However, some results show that the alignment quality of Dice coefficient is lower than the statistical models [1]. Och gained the alignment error rate for Dice model in the best case something about 30 percent. But they demonstrated statistical models outperform the simple Dice algorithm.

4. INITIAL EXPERIMENTS

We started our alignment experiment with implementation of Expectation Maximization and Dice Coefficient algorithms and tested them with one hundred training data. Once the training process was completed, a simple manual evaluation was conducted using a manual prepared reference for 20 sentence pairs. The reference contained the correct alignment of those sentence pairs. However, since our training data are not in an adequate amount, the evaluation process is not complete yet.

The following Maori-English sentences are a typical input for both Dice and EM algorithms. The results for both algorithms are shown in Table 1 and Table 2 respectively. The correct alignments are marked as highlighted.

Maori: *Ko te mihi tuatahi ki te kaihanga.*

English: *My first acknowledgement goes to our heavenly father*.

Table 1. Dice alignment output for a sample sentence pair. Highlighted entries are correct ones.

Maori	English
ko	our
te	to
mihi	acknowledgement
tuatahi	first
ki	to
te	to
kaihanga	goes

Table 2. Expectation Maximization alignment for a sample sentence pair

Maori	English
ko	to
te	my
mihi	acknowledgement
tuatahi	first
ki	our
te	goes
kaihanga	heavenly

Although we have not done a thorough evaluation over the results of two alignment methods yet, exploring the initial results show that Dice Coefficient has done the alignment better that the other algorithm with current tiny training data. Despite having noisy data, the generated alignments have produced some alignments completely or partially correct.

5. CONCLUSION

Maori is among the languages with limited online resources. For the aim of statistical machine translation, huge amount of language resources is critical.

In this research, we intended to develop a basic component of Statistical Machine Translation for Maori-English language pair. The results gained by a small size of data encouraged us to expand our work towards this goal. Both alignment algorithms (Expectation Maximization and Dice) used in this research were able to produce some complete or partial correct alignments. We intend to enrich our parallel corpus by gathering more bilingual texts from online resources as well as manual translations. We also intend to improve and expand our alignment methods to gain a higher accuracy in the phrase lookup table.

6. REFERENCES

[1] Och, F. J. 2002. Statistical Machine Translation: From Single-Word Models to Alignment Templates, Ph.D thesis.

[2] http://www.statmt.org/survey/

[3] Och, F. J. and Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, 29(1):19-51

[4] Lopez, A. and Resnik, P. 2006. Word-Based Alignment, Phrase-Based Translation: What's the Link?, Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 90-99, Cambridge.

[5] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263–311.

[6] Vogel, S., Ney, H., and Tillmann, C. 1996. HMM-based Word Alignment in Statistical Translation. In COLING '96: The 16th International Conference on Computational Linguistics, pp. 836-841, Copenhagen, Denmark.

[7] Jorg, T. 2004. Word to word alignment strategies, COLING 2004.

[8] Chiang, D. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*

[9] Tiedemann, J. 2003. Recycling Translations: Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing. Acta Universitatis Upsaliensis: Studia linguistic upsaliensia, ISSN 1652-1366, ISBN 91-554-5815-7.

[10] Charitakis, K. 2007. Using parallel corpora to create a Greek-English dictionary with Uplug, in Proc. 16th Nordic Conference on Computational Linguistics -NODALIDA '07.

[11] Megyesi, B. and Dahlqvist, B. 2007. The Swedish-Turkish Parallel Corpus and Tools for its Creation, in Proc. of 16th Nordic Conference on Computational Linguistics - NODALIDA '07.

[12] Velupillai, S. and Dalianis, H. 2008. Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic Languages, Coling 2008: Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, pages 10–16, Manchester

[13] Xing, H., Zhang X. 2008. Using parallel corpora and Uplug to create a Chinese-English dictionary, Master Thesis, Department of Computer and Systems Sciences, KTH/Stockholm University, Sweden.

[14] Gao, Q., Bach, N., and Vogel, S. 2010. A semi-supervised word alignment algorithm with partial manual alignments. In In Proceedings of the ACL 2010 joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (ACL-2010 WMT).

[15] Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1):1–38.

[16] Mermer, C. and Saraclar, M. 2011. Bayesian Word Alignment for Statistical Machine Translation, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers, pages 182–187, Portland, Oregon, June 19-24.

[17] Knight, K. and Koehn, P. 2003. What's new in statistical machine translation, University of South California,

[18] Melamed, I. D. 2000. Models of translational equivalence among words. *Computational Linguistics*, Vol. 26, No. 2, pp. 221–249.